Biostatistics (2005), **6**, 2, *pp.* 227–239
doi:10.1093/biostatistics/kxi005

On criteria for evaluating models of absolute risk

MITCHELL H. GAIL*, RUTH M. PFEIFFER

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Executive Plaza South, EPS 8032, Bethesda, MD 20892-7244, USA
gailm@mail.nih.gov

SUMMARY

Absolute risk is the probability that an individual who is free of a given disease at an initial age, a, will develop that disease in the subsequent interval $(a,t]$. Absolute risk is reduced by mortality from competing risks. Models of absolute risk that depend on covariates have been used to design intervention studies, to counsel patients regarding their risks of disease and to inform clinical decisions, such as whether or not to take tamoxifen to prevent breast cancer. Several general criteria have been used to evaluate models of absolute risk, including how well the model predicts the observed numbers of events in subsets of the population ('calibration'), and 'discriminatory power,' measured by the concordance statistic. In this paper we review some general criteria and develop specific loss function-based criteria for two applications, namely whether or not to screen a population to select subjects for further evaluation or treatment and whether or not to use a preventive intervention that has both beneficial and adverse effects. We find that high discriminatory power is much more crucial in the screening application than in the preventive intervention application. These examples indicate that the usefulness of a general criterion such as concordance depends on the application, and that using specific loss functions can lead to more appropriate assessments.

Keywords: Absolute risk model; Accuracy; Loss functions for clinical decisions; Negative predictive value; Positive predictive value; ROC curve.

1. INTRODUCTION

Absolute risk, also called 'crude probability' (e.g. Tsiatis, 1998), is the probability that a subject who is free of the disease of interest at age a will be diagnosed with that disease in a subsequent age interval $(a,t]$. To be explicit, if $h_1(u)$ is the cause-specific hazard at age u for the disease of interest and $h_2(u)$ is the hazard of mortality from other causes, the absolute risk is

$$\pi = \int_a^t h_1(u) \exp\left[-\int_a^u \{h_1(v) + h_2(v)\}\, \mathrm{d}v\right] \mathrm{d}u. \tag{1.1}$$

Covariates are often used to model h_1 and yield covariate-specific estimates of absolute risk that have several uses. For example the breast cancer risk model of Gail *et al*. (1989), as modified (Anderson *et al*., 1992; Costantino *et al*., 1999), has been used to plan intervention trials (Fisher *et al*., 1998), to inform specific clinical decisions, such as whether or not to take tamoxifen to prevent breast cancer (Gail *et al*., 1999), and to assist women who come for counseling.

*To whom correspondence should be addressed.

Several books (e.g. Hand, 1981, 1997; McLachlan, 1992; Pepe, 2003) and articles (e.g. Hand, 2001) describe general methods to evaluate models for dichotomous outcomes, such as models of absolute risk. In the medical literature, a variety of general criteria have been used to assess risk models. For example independent validation studies have shown that the modified Gail model is 'well calibrated' in the sense that the expected numbers (E) of breast cancers predicted for subsets of women agree well with the observed numbers (O) of cancers that actually develop (Costantino *et al*., 1999; Rockhill *et al*., 2001). Goodness-of-fit criteria based on O and E, including examination of the ratio O/E in the overall population and in subsets, have been used to assess unbiasedness (or 'calibration'). This model has been criticized, however, on the grounds that it cannot reliably discriminate women who will develop breast cancer from those who will not (Rockhill *et al*., 2001). One commonly used measure of discrimination is the concordance statistic, c, which is the probability that a randomly selected woman with breast cancer will have a higher projected risk than a randomly selected woman without breast cancer.

In this article, we present a general framework for assessing models of absolute risk and review widely used general criteria (Section 2). We then contrast these general criteria with criteria based on loss functions that are tailored to specific applications. We introduce a decision theoretic framework for determining whether a preliminary population screening instrument is useful for selecting individuals who require more definitive diagnostic evaluation or intervention (Section 3), and for deciding whether or not to recommend an intervention that affects one health outcome favorably and another unfavorably (Section 4). We show that the losses from using a model with imperfect discriminatory power are greater in the screening application than in the intervention application. We illustrate these ideas further with an example based on the Gail model (Section 5) before drawing conclusions in Section 6.

2. NOTATION AND GENERAL CRITERIA FOR ASSESSING ABSOLUTE RISK MODELS

We assume a sample of N individuals from an infinite population. We are interested in whether the i-th individual will be diagnosed with a particular disease in the next 5 years ($Y_i = 1$) or not ($Y_i = 0$). Individual i has a true probability $\pi_i = P(Y_i = 1)$ defined as in (1.1) and a vector of predictors X_i, and we consider a risk prediction model $r(x)$, which is a mapping from the set, Ω, of possible values of X to [0,1]. The quantity π_i contains all the information about the risk of individual i. The joint distribution of (Y, π, X) is $\pi^Y(1-\pi)^{1-Y}G(\pi, X)$, where G is the joint distribution of π and X, and Y is conditionally independent of any other factors given π. The distribution G gives rise to the marginal distributions G_x and G_π and the conditional distribution $G(\pi|x)$. This formulation accommodates determinists, who assume π_i is 1 or 0, as well as those who allow $0 < \pi_i < 1$. Although we would like to know π_i for individual i, we typically have access only to a risk model prediction $r(x_i)$ based on covariates $X = x_i$. The model $r(x)$ is "perfectly calibrated" if

$$r(x) = \eta(x) \equiv E(\pi|x) = \int \pi \, \mathrm{d}G(\pi|x), \quad \text{for each } x. \tag{2.1}$$

Some risk models $r(x)$ map many values of x into a single risk. For example all x values within given ranges of a continuous distribution may be assigned the same risk. If $r(x)$ is such a many-to-one function, it induces a partition of Ω into subsets $A_1, A_2, \ldots, A_k$ on which r is constant. If x is any value in subset A_j, then r is "weakly calibrated" if $r(x) = E\{\pi|x \in A_j\}$ for all $j = 1, 2, \ldots, k$. A weakly calibrated model is perfectly calibrated if π is constant for every $x \in A_j$, for all $j = 1, 2, \ldots, k$. The distribution of r is

$$F(r) = \int_{\{x: r(x) \leqslant r\}} \mathrm{d}G_x(x). \tag{2.2}$$

If r is perfectly calibrated, $F(r)$ equals

$$F_\eta(r) \equiv P[\eta(x) \leqslant r] = \int_{\{x:\,\eta(x)\leqslant r\}} \mathrm{d}G_x(x), \tag{2.3}$$

but otherwise (2.2) and (2.3) need not be equal. Before we discuss a decision theoretic framework for assessing the performance of model $r(x)$, we review four classes of general criteria that are used for this purpose. We express these criteria in terms of $r(x_i)$ and $F(r)$, but they can be applied to π_i and $F_\eta(r)$. We will illustrate some of the criteria with two risk models chosen to describe the probability of a rare outcome, $P(Y_i = 1) = E(r) = \mu = 0.01$. The first is a Beta distribution Beta(0.01,0.99) with standard deviation 0.070; the second is Beta(1,99) with standard deviation 0.0099. Thus, the first reflects considerably more dispersed risk and a better opportunity to discriminate individuals who will develop disease from those who will not.

2.1 *Calibration*

For simplicity, suppose we have an 'external' sample $(Y_i, \pi_i, X_i)_{i=1}^N$ that is independent of any data used to estimate $r(x)$. Several measures of calibration or goodness-of-fit of the model $r(x)$ to the data are commonly recommended. One approach is to partition Ω into measurable sets $A_1, A_2, \ldots, A_L$ and compare the observed numbers of events $O_k = \sum Y_i I(X_i \in A_k)$ with the expected numbers of events $E_k = \left\{\sum r(X_i) I(X_i \in A_k)\right\}$. If $r(x) \ll 1$ for all X, then O_k is approximately Poisson, and if $r(x)$ is well calibrated, O_k has mean E_k. Thus, confidence intervals can be constructed for ratios such as O_k/E_k or its reciprocal (see e.g. Costantino *et al*., 1999; Rockhill *et al*., 2001). An L degree-of-freedom global goodness-of-fit statistic, $\sum (O_k - E_k)^2/E_k$ can also be computed. A commonly used partitioning is based on ordering $r(x)$ and grouping together all risk factors X that correspond to deciles of r (Lemeshow and Hosmer, 1982).

Another criterion in this class is the mean squared error, or Brier (1950) statistic, $B = N^{-1} \times \sum \{Y_i - r(X_i)\}^2$, which can be decomposed into bias and intrinsic variability,

$$B = N^{-1}\left[\sum \{\pi_i - r(X_i)\}^2 + \sum (Y_i - \pi_i)^2\right]. \tag{2.4}$$

Equation (2.4) demonstrates that even for a perfect model in which $r(X_i) = \pi_i$ for all i, there is inherent variability in the predicted outcome, $N^{-1} \sum (Y_i - \pi_i)^2 = \mathrm{Var}(Y_i)$, unless the process is deterministic with $\pi_i = 1$ or 0 for all i.

2.2 *Discrimination*

A second class of criteria describes how well the decision rule $r(x)$ "discriminates" those who will develop disease from those who will not. If r is perfectly calibrated, the distribution of r among cases ($Y = 1$) is

$$F_{\text{case}}(v) \equiv P(r \leqslant v | Y = 1) = \mu^{-1}\left\{\int_o^v r \,\mathrm{d}F(r)\right\}, \tag{2.5}$$

where $\mu = \int r \,\mathrm{d}F(r)$ is the mean risk in the general population. Likewise, the distribution of r among those with $Y = 0$ is

$$F_{\text{control}}(v) \equiv P(r \leqslant v | Y = 0) = (1 - \mu)^{-1} \int_0^v (1 - r) \,\mathrm{d}F(r). \tag{2.6}$$

Figure 1(a) displays F_{case} for Beta(0.01,0.99) (solid) and for Beta(1,99) (dash); corresponding distributions F_{control} are shown as dash–dots and dots, respectively. Only for Beta(0.01,0.99) are F_{case} and F_{control} well separated.

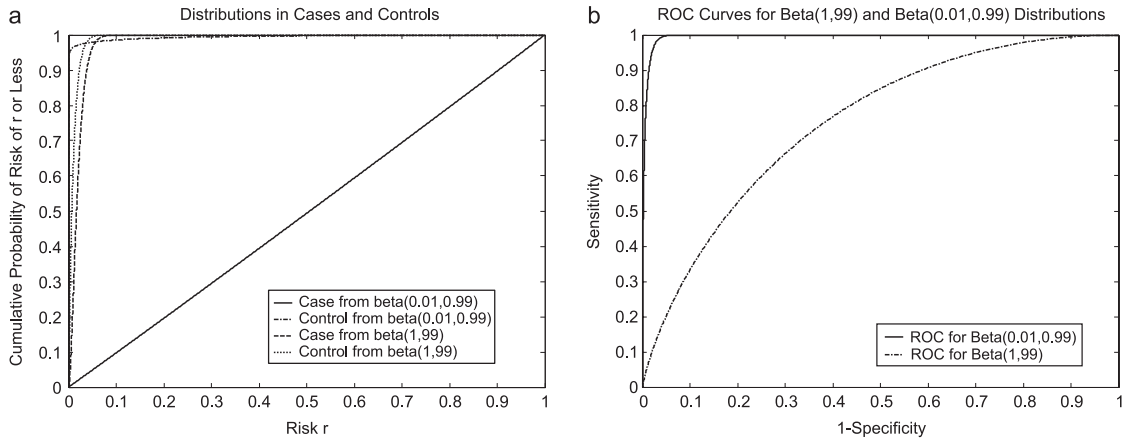

Fig. 1. (a) Distributions F_{case} and F_{control} for cases and controls from the underlying distributions $F = \text{Beta}(0.01, 0.99)$ and $F = \text{Beta}(1, 99)$. (b) ROC curves for the Beta(0.01,0.99) distribution (solid) and the Beta(1,99) distribution (dash–dot).

Note that F_{case} and F_{control} and any functionals of them can be estimated from random samples of cases and non-cases and do not require prospective follow-up, unlike calibration criteria. The sensitivity of a classification rule based on cutoff r^* is $\text{sens}(r^*) = P(r \geqslant r^* | Y = 1) = 1 - F_{\text{case}}(r^* -)$, where $r^* -$ denotes a value just smaller than r^*, to allow for a discrete distribution F_{case}. Similarly, the specificity is $\text{spec}(r^*) = P(r < r^* | Y = 0) = F_{\text{control}}(r^* -)$. The ROC curve is a plot of $\text{sens}(r^*)$ against $1 - \text{spec}(r^*)$ as r^* varies from 0 to ∞, as illustrated by the solid and dash–dotted loci in Figure 1(b) for Beta(0.01,0.99) and Beta(1,99), respectively. The area under the ROC curve, AUC, also called the "concordance," c, is the probability that a randomly selected r from F_{case} will exceed a randomly selected r from F_{control}. The AUCs are 0.995 and 0.751, respectively, for the underlying distributions Beta(0.01,0.99) and Beta(1,99). Several authors recommended the partial area under the ROC curve corresponding to ranges of specificity of medical interest as being preferable to AUC (Wieand *et al.*, 1989; Pepe, 2003; Dodd and Pepe, 2003).

Features of F itself give insight into the discriminatory power of $r(x)$. For example if values of $r(x)$ are near 0 and 1 only, then F_{case} will be concentrated near 1 and F_{control} near 0. Various measures of dispersion of F can be used to characterize discrimination. The Lorenz curve (e.g. Dagum, 1985),

$$L(p) = \mu^{-1} \int_0^{\xi_p} r \, \mathrm{d}F(r), \tag{2.7}$$

describes the proportion of population risk that falls at or below the p-th quantile of risk, $\xi_p = F^{-1}(p)$, as shown in Figure 2 for the Beta(0.01,0.99) and Beta(1,99) distributions. The proportion of the population risk in the $100p\%$ of the population at highest risk is $1 - L(1 - p)$. If risk is highly concentrated, the model $r(x)$ can be useful for detecting a small proportion of the population that carries most of the risk and requires preventive intervention (Pharoah *et al.*, 2002). For example for $p = 0.1$, $1 - L(0.9)$ is 1.000 for Beta(0.01,0.99) and only 0.103 for Beta(1,99). The Gini index is twice the area between the equiangular line (a plot of p against p) and the Lorenz curve (a plot of $L(p)$ against p for $0 \leqslant p \leqslant 1$). The Gini index is a measure of risk inequality in the population. The Gini indices are, respectively, 0.980 and 0.498 for the Beta(0.01,0.99) and Beta(1,99) distributions. Sensitivity at $r^* = \xi_p$ is $1 - L(p)$. For rare diseases with $r(x) \ll 1$, specificity at $r^* = \xi_p$ is approximately p. Hence, the AUC approximately equals the

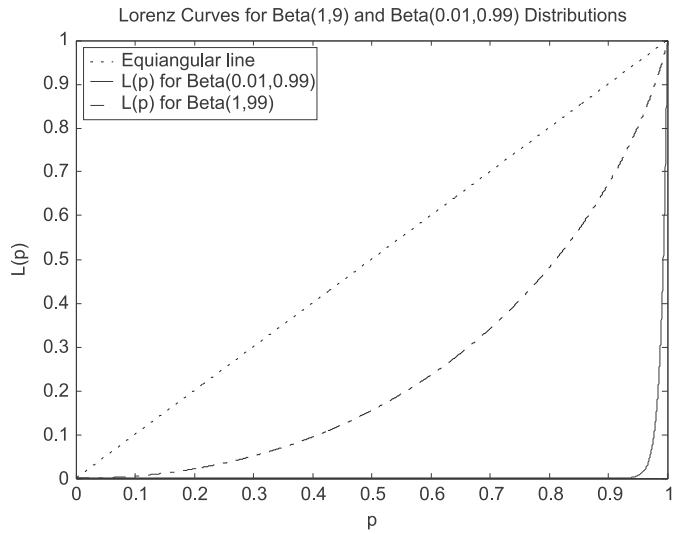

Fig. 2. Lorenz curves for Beta(1,99) distribution (dash–dot) and Beta(0.01,0.99) distribution (solid). The Gini index is twice the area between the equiangular line (dotted) and the Lorenz curve.

area above the Lorenz curve, and twice the AUC minus one approximates the Gini index. The values for $2(\text{AUC}) - 1$ are 0.990 and 0.503, respectively, for the Beta(0.01,0.99) and Beta(1,99) distributions, in good agreement with the Gini indices because $P(Y = 1) = 0.01$ in these examples.

2.3 *Accuracy*

A third class of criteria, termed 'accuracy' by Hand (2001), refer to how accurately a dichotomization $r \geqslant r^*$ versus $r < r^*$ classifies the outcomes Y. Here r^* is called the 'cutoff' value. Two commonly used measures of accuracy are positive predictive value, $P(Y = 1|r \geqslant r^*)$ and negative predictive value, $P(Y = 0|r < r^*)$. The total correct classification probability, $P(r \geqslant r^*)P(Y = 1|r \geqslant r^*) + P(r < r^*)P(Y = 0|r < r^*)$, or its complement, the misclassification probability $P(r \geqslant r^*)P(Y = 0|r \geqslant r^*) + P(r < r^*)P(Y = 1|r < r^*)$, is sometimes taken as measures of accuracy or inaccuracy, respectively. These latter criteria would be appropriate for a decision problem in which each type of misclassification carried equal loss and no loss is incurred with correct classification. Public health decisions to screen a population (Section 3) or clinical decisions to intervene to prevent disease (Section 4) do not have such symmetric losses, however. For all values in the range $0.05 < r^* < 0.95$, the negative predictive value and total correct classification probability are near $(1 - \mu) = 0.99$ for both the Beta(0.01,0.99) and Beta(1,99) distributions. For a rare disease, these quantities depend very little on sensitivity. By contrast, at $r^* = 0.25$, for example, the positive predictive value is 0.54 for Beta(0.01,0.99) and 0.26 for Beta(1,99).

2.4 *Proportion of variation explained*

Criteria have been proposed (e.g. Efron, 1978; Korn and Simon, 1991; Liao and McGee, 2003) to determine how much variability of risk is explained by the model $r(x)$. The entropy function $H(r) = -r\ \ln(r) - (1 - r)\ \ln(1 - r)$ describes the uncertainty of prediction. For r near 0 or 1, $H(r)$ is near

0 and $H(0.5)$ is the maximum entropy. The 'null risk model' assigns each person in the population the average risk μ and has entropy $H_0 = H(\mu)$. The null entropy is 0.056 for Beta(0.001,0.99) and Beta(1,99). Suppose instead, we use a well calibrated model $r(x)$ and compute the average conditional entropy, $\bar{H} = \int H(r)\,\mathrm{d}F(r)$. The fractional reduction in entropy or proportion of variation explained is $(H_0 - \bar{H})/H_0$, which is the same as the fractional reduction in binary deviance considered by Liao and McGee (2003). For the Beta(0.01,0.99) and Beta(1,99) distributions, the fractional reductions in entropies are 0.71 and 0.076, respectively, because the former represents a much more informative distribution of risks. See Haberman (1982) and Gilula and Haberman (1995) for other measures of dispersion.

3. DECISION THEORETIC MODEL FOR SCREENING

Although general criteria for model assessment are attractive because they can be used regardless of the application, more focused criteria are preferable for specific applications if one is willing to posit a model for loss associated with classification errors. We consider a screening application of a risk model $r(x)$. For each subject, a decision is taken for further evaluation or preventive intervention, $\delta(x) = 1$, or no further evaluation or intervention, $\delta(x) = 0$, based on measured covariates $X = x$. Depending on the true but unknown eventual state of the individual, Y, the losses are as shown in Table 1. We assume that the screening operation has no impact on $\pi_i = P(Y_i = 1)$. For an arbitrary decision rule, $\delta(x)$, the expected loss is

$$\int [C_{11}\delta(x)\eta(x) + C_{10}\delta(x)\{1-\eta(x)\} + C_{01}\{1-\delta(x)\}\eta(x) + C_{00}\{1-\delta(x)\}\{1-\eta(x)\}]\,\mathrm{d}G_x, \quad (3.1)$$

which is minimized by minimizing the integrand for

$$\begin{cases} \delta(x) = 1, & \text{if } \eta(x) \leqslant (C_{10} - C_{00})/(C_{10} + C_{01} - C_{00} - C_{11}), \\ \delta(x) = 0, & \text{otherwise.} \end{cases} \quad (3.2)$$

For a perfectly calibrated model $r(x)$ with $\delta(x) = 1$ for $r \geqslant r^*$ and $\delta(x) = 0$ otherwise, the expected loss is

$$\mathrm{EL} = C_{11}\int_{r^*}^{1} r\,\mathrm{d}F(r) + C_{01}\int_{0}^{r^*} r\,\mathrm{d}F(r) + C_{10}\int_{r^*}^{1}(1-r)\,\mathrm{d}F(r) + C_{00}\int_{0}^{r^*}(1-r)\,\mathrm{d}F(r). \quad (3.3)$$

The value, r^*, that minimizes EL is

$$r^* = (C_{10} - C_{00})/(C_{10} + C_{01} - C_{00} - C_{11}). \quad (3.4)$$

The minimal expected loss at r^* is

$$\mathrm{EL}_{\min} = C_{11}\mu\,\mathrm{sens}(r^*) + C_{01}\mu\,\{1-\mathrm{sens}(r^*)\} + C_{10}(1-\mu)\{1-\mathrm{spec}(r^*)\} + C_{00}(1-\mu)\mathrm{spec}(r^*). \quad (3.5)$$

Table 1. *Losses for a screening problem*

Decision	Diseased, $Y = 1$	Not diseased, $Y = 0$
Further evaluation, $\delta = 1$	C_{11}	C_{10}
No further evaluation, $\delta = 0$	C_{01}	C_{00}

Equation (3.5) is equivalent to (2.17) in Pepe (2003) if $C_{00} = 0$. If $r(x)$ has perfect discrimination at r^*, $\text{sens}(r^*) = \text{spec}(r^*) = 1$, and

$$\text{EL}_{\text{perfect}} = C_{11}\mu + C_{00}(1-\mu). \tag{3.6}$$

To illustrate these ideas, consider a self-administered questionnaire designed to estimate risk of colorectal cancer in the next 5 years. The decision $\delta = 1$ denotes referring the subject for further evaluation, such as colonoscopy, whereas $\delta = 0$ denotes no further evaluation. Compared to the consequences of misclassifying a subject, we assume that the costs of administering and interpreting the questionnaire are negligible ($C_{00} = 0$). Because colonoscopy or other diagnostic evaluations entail small but definite risks, such as bleeding or perforation, we assign losses $C_{10} = 1$. The greatest cost is $C_{01} = 100$, because failing to evaluate a subject who will develop colon cancer in the next 5 years ($Y = 1$) can result in death or severe morbidity. The cost $C_{11} = 0.1C_{01} + C_{10} = 11$ is chosen on the basis of the fact that the risks of colonoscopy must still be borne ($C_{10} = 1$), but the early evaluation may reduce the risks of death or severe morbidity in a person otherwise destined to develop colon cancer by an appreciable factor, which we take to be 0.1. The critical risk is $r^* = 1/(1 + 100 - 11) = 1/90$, and $\text{EL}_{\text{perfect}} = \mu C_{11} = 0.1100$ for $\mu = 0.01$. For any other rule,

$$\text{EL}_{\text{min}} = 11\mu\,\text{sens}(r^*) + 100\mu\{1 - \text{sens}(r^*)\} + (1-\mu)\{1 - \text{spec}(r^*)\}. \tag{3.7}$$

One way to assess how much the loss is inflated by using $r(x)$ instead of a rule with perfect sensitivity and specificity is to compute the loss ratio $= \text{EL}_{\text{min}}/\text{EL}_{\text{perfect}}$. If all subjects are screened, sens = 1.0, spec = 0.0 and the expected loss is 1.10, a 10-fold increase over $\text{EL}_{\text{perfect}}$. If none are screened (sens = 0.0, spec = 1.0), the loss is 1.00, a 9.09-fold increase above the minimum risk. Figure 3 shows the loss ratio for various values of sensitivity and specificity. A test with sensitivity 0.9 and specificity 0.2 has loss ratio of 9.01. A test with sensitivity 0.2 and specificity 0.6 has loss ratio of 9.27. These are the kinds

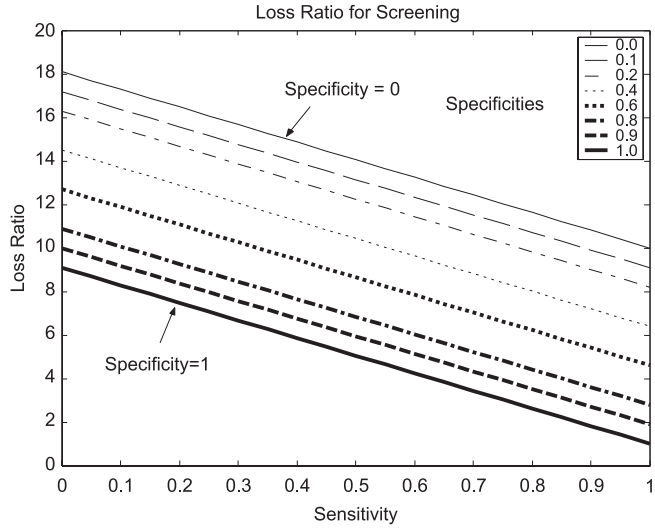

Fig. 3. Plot of loss ratio for screening against sensitivity for various values of specificity. Comparison is against a model with perfect sensitivity and specificity. Parameters defined in Section 3 include overall disease risk $\mu = 0.01$ and costs $C_{11} = 11$, $C_{10} = 1$, $C_{01} = 100$ and $C_{00} = 0$ (see Table 1).

of sensitivities and specificities that a test with modest discriminatory power could exhibit, as shown in Figures 1 and 2 of Pepe *et al*. (2004). Much higher sensitivities and specificities are needed to achieve a loss ratio near 1.0 for this screening example (Figure 3).

4. A MODEL FOR CLINICAL DECISION-MAKING

We present a decision theoretic framework for deciding whether or not to use a preventive intervention that affects some endpoints favorably and some unfavorably. For example Gail *et al*. (1999) weighed the benefits of tamoxifen in reducing risks from breast cancer and hip fracture against the increased risks from stroke, pulmonary emboli and endometrial cancer in women taking tamoxifen. For ease of exposition, we consider only two outcomes and use the terminology of invasive breast cancer and stroke, but the ideas extend to multiple outcomes.

We let $\delta(x) = 1$ or $\delta(x) = 0$ according to whether a woman is given a preventive intervention or not. Each woman has two outcome variables, $Y_1^\delta = 1$ or 0 according to whether she develops invasive breast cancer or not over a defined time interval, and $Y_2^\delta = 1$ or 0 is defined similarly for stroke. Although we only get to observe (Y_1^δ, Y_2^δ) for $\delta = 1$ or for $\delta = 0$ because each woman either receives the intervention or not, the complete counterfactual (e.g. Maldonado and Greenland, 2002) data describing outcomes in the presence and absence of intervention are $(Y_1^0, Y_2^0, Y_1^1, Y_2^1)$. The effect of intervention is to change the distribution from $P_0(Y_1^0, Y_2^0)$ to $P_1(Y_1^1, Y_2^1)$. We extend the notation in Section 2 to allow for two bivariate outcomes and intervention by defining the vectors $\pi_\delta = \{P_\delta(Y_1^\delta = 1, Y_2^\delta = 1), P_\delta(Y_1^\delta = 1, Y_2^\delta = 0), P_\delta(Y_1^\delta = 0, Y_2^\delta = 1), P_\delta(Y_1^\delta = 0, Y_2^\delta = 0)\}^\top$ for $\delta = 1$ or 0.

Likewise, we define risk models $r_\delta(x) = \{r_{11}^\delta(x), r_{10}^\delta(x), r_{01}^\delta(x), r_{00}^\delta(x)\}^\top$ to predict the quadrinomial outcomes for $\delta = 1$ or 0. The i-th individual in the sample has random variables $(Y_{1i}^0, Y_{2i}^0, Y_{1i}^1, Y_{2i}^1, \pi_{0i}, \pi_{1i}, X_i)_{i=1}^N$ with joint distribution $Q(Y_{1i}^0, Y_{2i}^0|\pi_{0i})Q(Y_{1i}^1, Y_{2i}^1|\pi_{1i})G(\pi_{0i}, \pi_{1i}, X_i)$, where $Q(Y_{1i}^\delta, Y_{2i}^\delta|\pi_{0i})$ is the quadrinomial for outcomes for δ, and $G(\pi_{0i}, \pi_{1i}, X_i)$ is the joint distribution of π_{0i}, π_{1i} and X_i. This model induces a joint distribution, F, for $(r_0(x), r_1(x))$ and marginal distributions F_0 and F_1 for $r_0(x)$and $r_1(x)$, respectively. A model is perfectly calibrated if each of the components of $r_0(x)$ and $r_1(x)$ satisfies the equivalent of (2.1).

A decision is taken on the basis of X whether to intervene ($\delta = 1$) or not ($\delta = 0$). The corresponding losses are in Table 2. We assume that the losses from each disease outcome are additive. For example the loss for $\delta = 1$, $Y_1^1 = 1$ and $Y_2^1 = 0$ is $(C_{111} + C_{210})$. The additivity assumption might not be realistic in some settings, but our methods can be generalized for non-additivity by including separate losses for each combination of δ, Y_1^δ and Y_2^δ.

The additivity assumption simplifies the calculation of the expected loss. If the risk models are well calibrated, the expected loss for a decision rule $\delta(x) = 1$ or 0 is

$$\mathrm{EL}(\delta) = \int \left\{ \sum_{j=0}^{1} \sum_{k=0}^{1} (C_{11j} + C_{21k})\delta(x) r_{jk}^1(x) + (C_{10j} + C_{20k})(1 - \delta(x)) r_{jk}^0(x) \right\} \mathrm{d}G_x(x), \qquad (4.1)$$

Table 2. *Losses for each clinical outcome associated with a preventive intervention* ($\delta = 1$) *or non-intervention* ($\delta = 0$)

Intervention status	Breast cancer		Stroke	
	$Y_1^\delta = 0$	$Y_1^\delta = 1$	$Y_2^\delta = 0$	$Y_2^\delta = 1$
$\delta = 1$	C_{110}	C_{111}	C_{210}	C_{211}
$\delta = 0$	C_{100}	C_{101}	C_{200}	C_{201}

where G_x is the marginal distribution of X. Minimizing the integrand in (4.1) for each x leads to the optimal decision rule $\delta^*(x)$ given by

$$\begin{cases} \delta^*(x) = 1, & \text{if } \sum_{j=0}^{1} \sum_{k=0}^{1} \left\{ (C_{11j} + C_{21k}) r_{jk}^1(x) - (C_{10j} + C_{20k}) r_{jk}^0(x) \right\} \leqslant 0, \\ \delta^*(x) = 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

The corresponding minimal loss for the risk models $r_0(x)$and $r_1(x)$ is

$$\mathrm{EL}_{\min} = \mathrm{EL}(\delta^*). \tag{4.3}$$

If we had complete information $(Y_1^0, Y_2^0, Y_1^1, Y_2^1)$, we would be able to choose the optimal strategy for each woman by intervening only if $C_{11Y_{1i}^1} + C_{21Y_{2i}^1} \leqslant C_{10Y_{1i}^0} + C_{20Y_{2i}^0}$. Thus, the expected loss with this complete information would be

$$\mathrm{EL}_{\text{perfect}} = \iiint \min(C_{11Y_1^1} + C_{21Y_2^1}, C_{10Y_1^0} + C_{20Y_2^0})\, \mathrm{d}Q(Y_1^0, Y_2^0 | \pi_0)\, \mathrm{d}Q(Y_1^1, Y_2^1 | \pi_1)\, \mathrm{d}G(\pi_0, \pi_1), \tag{4.4}$$

where $G(\pi_0, \pi_1)$ is the indicated marginal distribution of $G(\pi_0, \pi_1, X)$. Equation (4.4) is analogous to (3.6) for a single outcome, but requires much more knowledge of the underlying probability structure. Moreover, with data available only on (Y_1^0, Y_2^0)or on (Y_1^1, Y_2^1) for each subject, one cannot estimate the joint distribution$G(\pi_0, \pi_1)$ needed to compute (4.4).

Although one cannot compute $\mathrm{EL}_{\text{perfect}}$ from (4.4), one can still hope to obtain a good decision rule from (4.2) and estimate its expected loss from (4.3). Two principal difficulties in implementing this approach are defining the losses in Table 2 and developing well calibrated joint risk models, $r_0(x)$and $r_1(x)$. Gail *et al.* (1999) considered age- and race-specific risks over 5-year age intervals and made assumptions under which one can carry out such an analysis. They assumed that invasive breast cancer and stroke carried equivalent losses that were much greater than the inconveniences and other losses of intervention. Under this assumption, one can set $C_{100} = C_{110} = C_{200} = C_{210} = 0$ and $C_{101} = C_{111} = C_{201} = C_{211} = 1$. Because both invasive breast cancer and stroke are rare events on a 5-year risk interval, a risk model might reasonably set $r_{11}^0(x) = 0$ and $r_{11}^1(x) = 0$. Because no risk factors were available to model stroke risk, Gail *et al.* used the average risk, s, to model $r_{01}^0(x) = s$ and $r_{01}^1(x) = \rho_s r_{01}^0(x) = \rho_s s$, where $\rho_s = 1.6$ is the factor by which treatment increases stroke risk (Fisher *et al.*, 1998). Similarly, $r_{10}^0(x) = b(x)$, where $b(x)$ is a well calibrated model for absolute invasive breast cancer risk, such as model 2 in Costantino *et al.* (1999). In the presence of tamoxifen, the risk of breast cancer is reduced (Fisher *et al.*, 1998) to approximately $r_{10}^1(x) = \rho_b r_{10}^0(x) = \rho_b b(x)$, where $\rho_b = 0.5$. The optimal rule is, from (4.2), $\delta^*(x) = 1$ if $b(x)(\rho_b - 1) + s(\rho_s - 1) \leqslant 0$ and $\delta^*(x) = 0$ otherwise.

One should intervene if $b(x) \geqslant (0.6/0.5)s = 1.2s$. The expected loss from (4.1) is

$$\mathrm{EL}_{\min} = \int \{ I(b(x) \geqslant 1.2s)(\rho_b b(x) + \rho_s s) + I(b(x) < 1.2s)(b(x) + s) \}\, \mathrm{d}G_x(x)$$

or, equivalently,

$$\mathrm{EL}_{\min} = \rho_b \mu_b \mathrm{sens}_b(1.2s) + \rho_s s\{1 - F_b(1.2s)\} + \mu_b\{1 - \mathrm{sens}_b(1.2s)\} + s F_b(1.2s). \tag{4.5}$$

In (4.5), μ_b is the mean breast cancer risk, $\mathrm{sens}_b(b^*)$is the sensitivity of the breast cancer risk model at $b(x) = b^*$ and F_b denotes the distribution of $b(x)$. For a 5-year risk projection, $b(x) \ll 1$ and $\mathrm{spec}_b(b^*) \equiv (1 - \mu_b)^{-1} \int_0^{b^*} (1 - b)\, \mathrm{d}F_b$. Hence, (4.5) is approximately

$$\mathrm{EL}_{\min} = \rho_b \mu_b \mathrm{sens}_b(1.2s) + \rho_s s\{1 - \mathrm{spec}_b(1.2s)\} + \mu_b\{1 - \mathrm{sens}_b(1.2s)\} + s\{\mathrm{spec}_b(1.2s)\}. \tag{4.6}$$

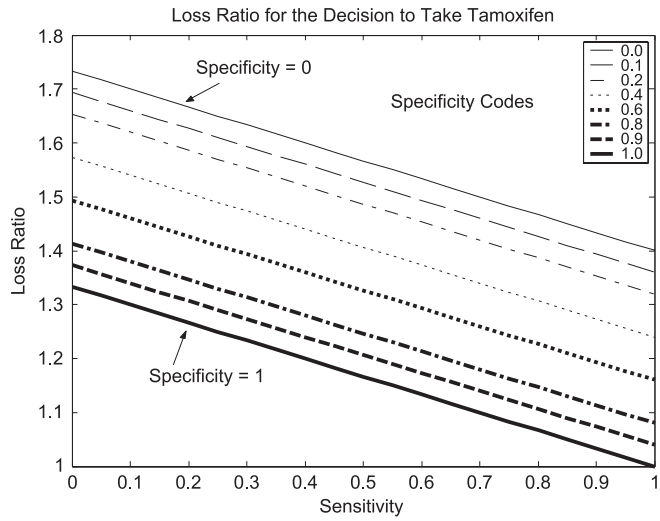

Fig. 4. Loss ratio for the clinical decision to give tamoxifen plotted against sensitivity of the model for breast cancer for various values of specificity. Comparison is against a breast cancer model with perfect sensitivity and specificity. Parameters in Section 4 include equal overall probabilities of breast cancer and stroke of 0.03, multiplicative effects of tamoxifen on chances of breast cancer of 0.5 and on stroke of 1.6 and losses (Table 2) of $C_{100} = C_{110} = C_{200} = C_{210} = 0$ and $C_{101} = C_{111} = C_{201} = C_{211} = 1$.

For $\mu_b = s = 0.03$, $\rho_s = 1.6$ and $\rho_b = 0.5$, (4.2) reduces to

$$\mathrm{EL}_{\min} = 0.078 - 0.015\,\mathrm{sens}_b(1.2s) - 0.018\,\mathrm{spec}_b(1.2s). \tag{4.7}$$

Although we cannot compare $\mathrm{EL}_{\min}$ for this joint risk model with $\mathrm{EL}_{\mathrm{perfect}}$ without making strong assumptions on $G(\pi_0, \pi_1)$, we can ask how well the given joint risk model works relative to a similar model with perfect sensitivity and specificity for the breast cancer diagnosis at $b^* = 1.2s$. The loss ratio compared to this model is (4.7) divided by $0.078 - 0.015 - 0.018 = 0.045$ (Figure 4). Loss ratios get no higher than 1.73, even for an implausibly bad model for breast cancer risk with zero sensitivity and specificity. For sensitivity 0.9 and specificity 0.2, the loss ratio is only 1.35, which can be compared to the loss ratio 9.01 for the screening application (Section 3). For sensitivity 0.2 and specificity 0.6, the loss ratio is only 1.43, which is much smaller than the loss ratio of 9.27 in the screening application (Section 3). Thus, improvements in the sensitivity and specificity of the breast cancer risk prediction do not yield dramatic reductions in expected loss concerning the decision to take tamoxifen.

5. EXAMPLE

To further illustrate the points in Sections 3 and 4, we analyzed data from the 2000 National Health Interview Survey (see Freedman *et al.*, 2003) to determine the distribution of 5-year absolute risk in women aged 55–60 years, as estimated from the modified Gail model (model 2 in Costantino *et al.*, 1999). The distribution had mean 0.0132 and standard deviation 0.0778. We approximated this distribution with a Beta(2.8386,211.83) distribution, which also has mean 0.0132 and standard deviation 0.0778. The corresponding AUC was 0.66. Compared to a model with perfect sensitivity and specificity, the loss ratio for the screening application in Section 3 was, from (3.7) with $r^* = 1/90$ and $\mu = 0.0132, 0.9589/0.1452 = 6.60$. Thus, based on the loss parameters in Section 3, the model would

not perform nearly as well as a model with perfect sensitivity and specificity for screening a population. This same model performed reasonably well for helping decide whether to take tamoxifen to prevent breast cancer based on the losses in Section 4. From data in Gail *et al*. (1999), we set the 5-year stroke risk for a 55-year-old woman to be $s = 0.0055$. Setting $\mu_b = 0.0132$, $\rho_b = 0.5$ and $\rho_s = 1.6$, we obtained $b^* = 1.2s = 0.0066$, $\text{sens}(b^*) = 0.933$ and $\text{spec}(b^*) = 0.200$. The loss ratio, compared to a breast cancer model with $\text{sens}(b^*) = \text{spec}(b^*) = 1$ was, from (4.6), 0.01518/0.121 = 1.25. Thus, the model performed reasonably well in this context, and there was comparatively little to be gained by improving the breast cancer risk model. Perhaps better performance could be obtained by refining the model for stroke risk instead.

6. Discussion

We presented a probabilistic framework for evaluating models of absolute risk, reviewed some general criteria for assessing such models (Section 2) and discussed a decision theoretic approach for screening a population for disease (Section 3) and for deciding whether to administer a preventive intervention that has adverse and beneficial effects (Section 4). Although general criteria are appealing because they can be compared across a variety of applications, their casual use can result in the choice of inferior criteria for assessing models in particular applications. Specifically, we believe there has been over-emphasis on the concordance statistic (c or AUC) as a summary measure of model performance, irrespective of the intended application. The examples in Sections 3, 4 and 5 indicate that high discriminatory power can be more important in a screening application than in deciding whether or not to take a preventive intervention that has both beneficial and adverse effects. Some of the popularity of the concordance statistic may be due to the fact that one can estimate it from samples of cases and controls, whereas estimation of criteria relating to calibration (Section 2.1), accuracy (Section 2.3), proportion of variation explained (Section 2.4) and expected loss (Sections 3 and 4) requires follow-up information from a cohort.

The decision theoretic approach to evaluation of diagnostic tests has been discussed by Pepe (2003, pp. 31–33, 71–72 and 269–275), and more general aspects of medical decision-making are described in Weinstein *et al*. (1980). Baker (2000) used the utility function, benefit x sensitivity $-$ cost x $(1 -$ specificity), to evaluate acceptable regions of sensitivity and specificity for screening applications and concluded that levels of specificity above 0.98 were often required, in agreement with the analysis in Section 3 indicating a need for high sensitivity and specificity.

One drawback of application-specific decision theoretic approaches is that it is difficult to describe and estimate realistic loss functions, leaving such analyses open to criticisms of subjectivity or of naive formulation of the true loss structure. It should be noted, however, that general criteria often reflect an implicit preference for a specific loss structure. For example the total misclassification rate is an optimal criterion if correct classification carries no loss and all types of misclassification carry equal loss. Likewise, various measures of the proportion of variability explained can be derived from particular loss functions (Gilula and Haberman, 1995).

The approach in this paper can be used to evaluate the effect on minimal expected loss that occurs when $r(x)$ is biased or is only weakly calibrated (Section 2). Additional work might include extensions to non-additive loss functions, inference for comparing two or more models with respect to these criteria and use of Bayesian methods to exploit the hierarchical structure in Section 4.

Acknowledgments

We thank Dr. Andrew N. Freedman for helpful discussions on evaluating risk models, Professor Joseph L. Gastwirth for useful comments on the Lorenz curve and the reviewer and editors.

REFERENCES

ANDERSON, S. J., AHNN, S. AND DUFF, K. (1992). *NSABP Breast Cancer Prevention Trial Risk. Assessment Program, Version 2*. Pittsburgh, PA: Department of Biostatistics, University of Pittsburgh.

BAKER, S. G. (2000). Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics* **56**, 1082–1087.

BRIER, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **75**, 1–3.

COSTANTINO, J. P., GAIL, M. H., PEE, D., ANDERSON, S., REDMOND, C. K., BENICHOU, J. AND WIEAND, H. S. (1999). Validation studies for models projecting the risk of invasive and total breast cancer incidence. *Journal of the National Cancer Institute* **91**, 1541–1548.

DAGUM, C. (1985). Lorenz curve. In Kotz, S. and Johnson, N. L. (eds), *Encyclopedia of Statistical Sciences*, Volume 5. New York: Wiley, pp. 156–161.

DODD, L. E. AND PEPE, M. S. (2003). Partial AUC estimation and regression. *Biometrics* **59**, 614–623.

EFRON, B. (1978). Regression and ANOVA with zero-one data: measures of residual variation. *Journal of the American Statistical Association* **73**, 113–121.

FISHER, B., COSTANTINO, J. P., WICKERHAM, D. L., REDMOND, C. K., KAVANAH, M., CRONIN, W. M., VOGEL, V., ROBIDOUX, A., DIMITROV, N., ATKINS, J., DALY, M., WIEAND, S., TAN-CHIU, E., FORD, L. AND WOLMARK, N. (1998). Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *Journal of the National Cancer Institute* **90**, 1371–1388.

FREEDMAN, A. N., GRAUBARD, B. I., RAO, S. R., MCCASKELL-STEVENS, W., BALLARD-BARBASH, R. AND GAIL, M. H. (2003). Estimates of the number of U.S. women who could benefit from tamoxifen for breast cancer chemoprevention. *Journal of the National Cancer Institute* **95**, 526–532.

GAIL, M. H., BRINTON, L. A., BYAR, D. P., CORLE, D. K., GREEN, S. B., SCHAIRER, C. AND MULVIHILL, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* **81**, 1879–1886.

GAIL, M. H., COSTANTINO, J. P., BRYANT, J., CROYLE, R., FREEDMAN, L., HELZLSOUER, K. AND VOGEL, V. (1999). Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. *Journal of the National Cancer Institute* **91**, 1829–1846.

GILULA, Z. AND HABERMAN, S. J. (1995). Dispersion of categorical variables and penalty functions: derivation, estimation, and comparability. *Journal of the American Statistical Association* **90**, 1447–1452.

HABERMAN, S. J. (1982). Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association* **77**, 568–580.

HAND, D. J. (1981). *Discrimination and Classification*. Chichester: John Wiley.

HAND, D. J. (1997). *Construction and Assessment of Classification Rules*. Chichester: John Wiley.

HAND, D. J. (2001). Measuring diagnostic accuracy and statistical prediction rules. *Statistica Neerlandia* **55**, 3–16.

KORN, E. L. AND SIMON, R. (1991). Explained residual variation, explained risk, and goodness of fit. *American Statistician* **45**, 201–206.

LEMESHOW, S. AND HOSMER, JR, D. W. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology* **115**, 92–106.

LIAO, J. G. AND MCGEE, D. (2003). Adjusted coefficients of determination for logistic regression. *The American Statistician* **57**, 161–165.

MALDONADO, G. AND GREENLAND, S. (2002). Estimating causal effects. *International Journal of Epidemiology* **31**, 422–429.

MCLACHLAN, G. J. (1992). *Discrimination Analysis and Statistical Pattern Recognition*. New York: John Wiley.

PEPE, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford: Oxford University Press.

PEPE, M. S., JANES, H., LONGTON, G., LEISENRING, W. AND NEWCOMB, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* **159**, 882–890.

PHAROAH, P. D., ANTONIOU, A., BOBROW, M., ZIMMERN, R. L., EASTON, D. F. AND PONDER, B. A. (2002). Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genetics* **31**, 33–36.

ROCKHILL, B., SPIEGELMAN, D., BYRNE, C., HUNTER, D. J. AND COLDITZ, G. A. (2001). Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. *Journal of the National Cancer Institute* **93**, 358–366.

TSIATIS, A. A. (1998). Competing risks. In Armitage, P. and Colton, T. (eds), *Encyclopedia of Biostatistics*, Volume I. Chichester: John Wiley, pp. 824–834.

WEINSTEIN, M. C., FINEBERG, H. V., ELSTEIN, A. S., FRAZIER, H. S., NEUHAUSER, D., NEUTRA, R. R. AND MCNEIL, B. J. (1980). *Clinical Decision Analysis*. Philadelphia, PA: W.B. Saunders.

WIEAND, S., GAIL, M. H., JAMES, B. R. AND JAMES, K. L. (1989). A family of nonparametric statistics for comparing diagnostic tests with paired or unpaired data. *Biometrika* **76**, 585–592.

[*Received May 11, 2004; revised November 9, 2004; accepted for publication November 9, 2004*]